

Predicción del rendimiento académico en las nuevas titulaciones de grado de la EPS de la Universidad de Córdoba

C. Romero, A. Zafra, E. Gibaja, M. Luque, S. Ventura

Departamento de Informática y Análisis Numérico

Universidad de Córdoba

Campus de Rabanales, Edificio C2

14011 Córdoba

cromero@uco.es, azafra@uco.es, egibaja@uco.es, mluque@uco.es, sventura@uco.es

Resumen

En este artículo se describe una experiencia de innovación docente que tiene como finalidad predecir el rendimiento académico de los estudiantes de primer curso de grado universitario utilizando técnicas de minería de datos. La investigación se ha realizado utilizando información de los alumnos de tres asignaturas de introducción a la informática que se han impartido durante el primer cuatrimestre del curso 2010-11 en las distintas titulaciones de grado de la Escuela Politécnica Superior de la Universidad de Córdoba.

Summary

In this paper, we describe an educational innovation experience about the prediction of the students' academic performance during the first year of university degree by applying data mining techniques. The research has been carried out by using the information of students of three subjects about introduction to computer science during the first semester of 2010-11 academic year in the different degrees of the Higher Technical College of the University of Cordoba.

Palabras clave

Rendimiento académico, Minería de Datos, Predicción, Clasificación, Informática.

Introducción

La predicción del rendimiento académico universitario es una tarea ciertamente compleja debido a que se ve afectada por varios indicadores que se ha demostrado que influyen en la

determinación de si un sujeto ha tenido éxito o ha fracasado en sus estudios. Estos indicadores son:

- Éxito o concordancia entre el tiempo teórico previsto para cada unidad (asignatura, curso, ciclo o titulación) y el tiempo que emplea el alumno en superarla.
- Abandono de los estudios.
- No superación de asignaturas, repetición de cursos y retraso en la finalización de estudios.

En los últimos años, el número de alumnos españoles que no superan sus estudios universitarios se sitúa entre el 30% y el 50% (según el Centro de Investigación, Documentación y Evaluación del MEC). Debido a que este valor puede considerarse demasiado alto, desde la administración se insiste en la necesidad de desarrollar mecanismos para combatir este fenómeno. En el marco del Espacio Europeo de Educación Superior, muchas universidades están teniendo en cuenta en sus planes estratégicos un objetivo primordial: reducir las tasas de abandono y reprobación de los alumnos, mediante no sólo un cambio metodológico, sino en los procesos de orientación, transición, adaptación y promoción de los estudiantes. Un claro ejemplo de este objetivo es que en las memorias de los títulos de grado, existe un apartado específico denominado *Resultados Previstos* donde se presenta una estimación realista de las tasas de graduación, abandono y eficiencia y el progreso de los alumnos de la titulación. Estos valores son normalmente obtenidos a partir de los resultados de años anteriores y son los utilizados, tanto por los sistemas de gestión de la calidad de cada Universidad como por la ANECA, como uno de los aspectos más importantes a considerar en la evaluación de las titulaciones.

Asignatura	Área	Titulación
Introducción a la Programación	Ciencias de la Computación	Grado en Ingeniería Informática
Fundamentos de Informática	Ciencias de la Computación	Grado en Ingeniería Electrónica
Fundamentos de Informática	Ciencias de la Computación	Grado en Ingeniería Mecánica

Tabla 1. Asignaturas involucradas en la experiencia.

Una forma de ayudar a resolver este problema es intentar predecir a tiempo el alumno que se encuentra en situación de riesgo o peligro para intentar evitar su fracaso utilizando técnicas de minería de datos. En esta línea, este artículo muestra una experiencia docente que consiste en la aplicación de técnicas de minería de datos para obtener un modelo de predicción del éxito, fracaso y abandono de los estudiantes en la asignatura de informática del primer curso de las distintas titulaciones de la Escuela Politécnica Superior (EPS) de la Universidad de Córdoba (UCO) durante el curso 2010/11. Este modelo nos puede servir para ayudar al estudiante a evitar un posible fracaso y al profesor en la toma de decisiones que puedan impactar en los resultados finales de los alumnos.

Asignaturas y directrices docentes

Esta experiencia de innovación docente ha sido realizada por un grupo de profesores que imparten docencia en tres asignaturas de primer curso de informática de las nuevas titulaciones de grado de la EPS de la UCO (ver Tabla 1).

Las tres asignaturas, de carácter básico en su correspondiente titulación, pertenecen a la rama de Ingeniería y Arquitectura y tienen como objetivo principal sentar las primeras bases en la programación de ordenadores. La carga lectiva de estas asignaturas es de 6 créditos ECTS y su duración está limitada al primer cuatrimestre.

Los contenidos teóricos que se imparten en estas asignaturas son exactamente los mismos:

- Introducción a la informática
- Fundamentos de la programación
- Tipos de datos simples
- Operadores y expresiones
- Estructuras de control
- Entrada y salida estándar
- Tipos de datos compuestos o vectores
- Programación estructurada y modular

La parte práctica dedica sus esfuerzos a que los alumnos asimilen los conceptos explicados en teoría programando delante de un ordenador ejercicios sencillos en el lenguaje de programación C. No obstante, se han adaptado algunos ejemplos a cada ingeniería en particular.

La metodología didáctica que se sigue en estas asignaturas consta de una parte presencial y otra no presencial. En la primera, las horas de teoría se dedican a explicar los principales conceptos, acompañándolos de la resolución de casos prácticos; las horas de laboratorio están dedicadas a que los alumnos pongan en práctica los conocimientos adquiridos en las clases teóricas resolviendo, delante de un ordenador, ejercicios sencillos de programación; además, existen otras horas presenciales dedicadas al trabajo en grupo (exposiciones, tutorías colectivas, debates, etc.).

La evaluación de estas asignaturas sigue, también, un mismo patrón. Los alumnos, para poder superar la asignatura, tienen que realizar una serie de pruebas objetivas en las que se les evalúa sobre los conceptos teóricos visto en clase, además de realizar una serie de ejercicios prácticos donde se evalúan las capacidades obtenidas en la programación de ordenadores.

El material del que disponen los alumnos para seguir las asignaturas es básicamente el mismo, conjunto de transparencias con la explicación de los contenidos teóricos y cuaderno de problemas y ejercicios que deben realizar a lo largo del curso. Todo este material se encuentra disponible en la plataforma Moodle. Existe una asignatura virtual (dentro de Moodle) por cada asignatura presencial donde, además de los recursos, se han utilizado dos actividades principalmente: los foros para consulta y resolución de dudas y las tareas para el envío de las prácticas

Descripción de la experiencia

La experiencia ha consistido en las siguientes etapas (ver Figura 1):

- Identificación de las posibles variables que afectan al rendimiento académico. Se trata de analizar cuál es la información que vamos a recopilar para obtener nuestros modelos.
- Captura y recogida de la mayor cantidad posible de datos referentes a las anteriores variables utilizando diferentes fuentes de datos como: encuestas, información proporcionada por el profesor, e información proporcionada por sistemas de enseñanza *on-line*.

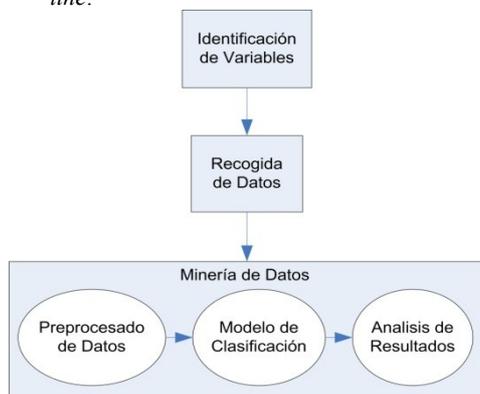


Figura 1. Proceso seguido en la experiencia.

- Aplicación de algoritmos de minería de datos sobre los datos anteriores. Para ello, se llevarán a cabo las siguientes etapas: preprocesado de los datos, creación de un modelo de clasificación/predicción, y análisis de los resultados y el modelo obtenido.

Identificación de variables

Debido a su gran repercusión, el problema de predecir el rendimiento académico, y más concretamente el fracaso y el abandono de los estudiantes se ha tratado ampliamente en todos los niveles educativos (enseñanza primaria, secundaria y superior). Este es un problema difícil de resolver, debido a la gran cantidad de factores que pueden afectar. De hecho es conocido como el problema de los cien factores .

Por ello, en esta primera fase se ha realizado una búsqueda bibliográfica sobre este

problema para el caso particular de enseñanzas universitarias . El objetivo es poder recopilar un conjunto de variables que ya han aparecido con frecuencia en artículos o trabajos anteriores sobre esta temática en concreto. En la Tabla 2, se muestra la lista de variables o factores (agrupados en tres categorías) que se han seleccionado de entre todos los factores encontrados.

Categorías	Variable/Factores
Personales	Vía de entrada a la universidad, notas anteriores, asistencia a clase, utilización de tutorías, interés por estudiar, expectativas de aprobar, nivel de motivación por aprobar, grado de satisfacción, tiempo de dedicación/estudio diario, técnicas de estudio, grado de interacción de la plataforma virtual de aprendizaje.
Profesorado, Pedagógicas e Institucionales	Grado de implicación del profesorado con el alumnado, grado de satisfacción del alumnado con las técnicas de evaluación, grado de dificultad o exigencia de la asignatura, grado de utilidad del conocimiento enseñado, calidad de los recursos docentes utilizados.
Familiares y Sociales	Nivel de apoyo de los padres al estudio, nivel económico de la familia, nivel de integración en clase.

Tabla 2. Variables o factores que pueden influir en el rendimiento académico.

A partir de las variables o factores concretos que se van a tener en cuenta en este estudio (ver Tabla 2) se determinaron las distintas fuentes de información o las formas de obtener dicha información.

Recogida y captura de datos

Durante la fase de recogida de datos se han utilizado diferentes fuentes de datos:

- Se ha recogido información sobre la nota y la vía u opción (prueba acceso universidad, ciclo formativo/formación profesional, simultaneo de estudios, titulado, discapacidad, extranjero o deportista de alto

nivel) de entrada de los alumnos a la Universidad de Córdoba en el curso 2010/11. Para ello se han consultado los datos proporcionados por la Junta de Andalucía sobre las distintas fases de adjudicación de plazas tanto en Junio como en Septiembre de 2010 para la EPS de la UCO.

- Se ha contabilizado la asistencia de los alumnos a las sesiones de prácticas de las asignaturas. Para ello se ha utilizado la hoja de firmas que pasa el profesor al comienzo de cada sesión de prácticas.
- Se ha realizado un seguimiento de la utilización y realización de actividades dentro del sistema de enseñanza virtual Moodle . Para ello se ha utilizado la información de *logs* o histórico de acciones realizadas por cada alumno dentro de la asignatura, ya que cada asignatura presencial tiene una asignatura virtual en Moodle con material y actividades. La información utilizada ha sido: número de actividades

realizadas, número total de accesos a Moodle, y número de recursos visitados. A partir de esta información se ha obtenido un único valor que nos mide el nivel de participación del alumno en la asignatura virtual y que tiene 3 posibles valores (alto, medio y bajo).

- Se ha elaborado una encuesta, que se ha pasado a los alumnos antes del final del cuatrimestre. La encuesta se ha realizado de forma electrónica como una actividad tipo test dentro del sistema Moodle . Concretamente, la encuesta utilizada consta de 19 preguntas (ver Tabla 3), cada una de ellas sobre alguna de variables seleccionadas (ver Tabla 2) para las que no se ha podido obtener de ninguna otra fuente de información. Para las posibles respuestas hemos utilizado una escala Likert, con 5 posibles opciones: Muy alto, Alto, Medio, Bajo y Muy bajo, donde el alumno debe de elegir una única opción.

#	Pregunta
1	Sobre la calidad de los recursos y materiales utilizados para impartir tanto las clases (aulas, laboratorios, ordenadores, etc.), ¿Qué nivel de calidad tienen para ti?
2	¿Qué nivel de dificultad y exigencia tiene para ti en general esta asignatura?
3	¿En qué nivel crees que los conocimientos que estas adquiriendo te serán luego de utilidad en tu vida profesional?
4	¿Con que grado has utilizado las tutorías y a los asesores académicos?
5	¿Cuál es tu grado de asistencia a las clases de teoría?
6	¿Qué ganas/interés tienes por estudiar esta asignatura en concreto y no otra?
7	¿Cuáles son tus expectativas de aprobar esta asignatura?
8	Con respecto a tus compañeros de clase. ¿Cual consideras que es tu nivel de integración?
9	¿Cuál es tu nivel de motivación por aprender lo que se enseña en la asignatura?
10	Con respecto a tu persistencia para aprobar esta asignatura a pesar de los obstáculos que te puedas encontrar, ¿Cuál es tu nivel de persistencia?
11	Las notas que has obtenido en tus anteriores estudios (bachillerato o ciclo formativo) antes de entrar a la Universidad, han sido en general:
12	Una vez que ya conoces mejor en qué consiste esta asignatura y que es lo que se enseña realmente: ¿Cuál es tu grado de satisfacción?
13	Con respecto a las técnicas de estudio como por ejemplo: subrayar los apuntes, realizar resúmenes y esquemas generales, lectura previa al comenzar a estudiar y repasar temas estudiados, etc. ¿En qué grado utilizas estas y otras técnicas de estudio?
14	Con respecto al tiempo que dedicas a la asignaturas, es decir, al trabajo personal diario adicional a la asistencia a clase. ¿Cuánto tiempo diario que le dedicas?
15	¿Qué nivel de apoyo, preocupación y motivación tienen tus padres ante tus estudios?
16	¿Qué nivel económico tiene tu familia?
17	Con respecto a si los profesores dialogan y motivan a los alumnos en la asignatura. En general, ¿Cuál crees que es el nivel de implicación de los profesores?
18	Sobre la forma de enseñar de los profesores, los métodos docentes utilizados y la calidad

8	educativa en las clases. ¿Cuál es tu grado de satisfacción?
1	Con respecto a las técnicas de evaluación que utilizan los profesores en las asignaturas:
9	exámenes, corrección de trabajos, exposiciones, etc. ¿Cuál es tu nivel de satisfacción?

Tabla 3. Enunciado de las preguntas de la encuesta pasada a los alumnos.

Asignatura	Total Alumnos	Tasa de aprobación	Tasa de reprobación	Tasa de abandono
Grado en Informática	123	56,09%	22,76%	21,13%
Grado en Electrónica	81	61,48%	19,87%	18,64%
Grado en Mecánica	76	60,58%	21,05%	18,42%

Tabla 4. Tasas de aprobación, reprobación y abandono en las asignaturas utilizadas en el estudio.

Finalmente, se han recogido las evaluaciones obtenidas por los alumnos en las asignaturas. Para ello se han utilizado las notas de las actas de junio y septiembre del 2011 publicadas por el profesor tras finalizar el cuatrimestre. Se ha utilizado tres valores categóricos: Aprobado, Suspenso, y No Presentado. En nuestro problema, este valor no es una variable o factor, sino que es la clase o valor a predecir.

Aplicación de Minería de Datos

Durante esta fase se aplicó un proceso típico de minería de datos, consistente en: pre-procesar los datos recogidos, construir y aplicar un modelo de minería de datos (un clasificador en nuestro caso), y finalmente analizar el modelo obtenido.

1.1. Pre-procesado de datos

Una vez recopilada toda la información, en primer lugar se convirtió a un mismo formato electrónico, ya que alguna venía en papel (el control de firmas), y otras en distintos formatos (la información de entrada en PDF, la información de Moodle en Excel, y las notas en TXT). Para ello, instalamos un sistema de gestión de bases de datos en un servidor, que también albergaba el software de minería de datos que se utilizó en la siguiente fase. De este modo, se integró toda la información anterior recogida sobre los alumnos en una única tabla resumen (con 24 variables o factores y una clase) donde cada fila o registro contiene toda la información disponible de cada alumno. A grandes rasgos, las variables que han sido contempladas son las siguientes:

- Nombre Alumno (variable 1)
- Información Entrada UCO (variable 2)
- Asistencia a prácticas (variable 3)

- Utilización de Moodle (variable 4)
- Encuesta (variables 5 a 24)
- Nota Final (clase a predecir)

Después, se realizaron una serie de operaciones para preparar los datos para el proceso de minería:

- Se eliminó la primera variable (nombre del alumno), para mantener la privacidad de la información de cada alumno en particular y no tener problemas legales con respecto a la Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal .
- Se eliminaron algunos alumnos. Concretamente, no se han tenido en cuenta a los alumnos que no realizaron la encuesta. Esto es debido a que la encuesta es la fuente de información más amplia de la que disponemos (19 variables de las 24 totales).

Finalmente, se dispone de un total de 280 alumnos que presentan unos porcentajes (tasas) de aprobación (éxito o aprobar), reprobación (fracaso o suspender) y abandono (no presentado) muy parecidos en las tres asignaturas (ver Tabla 4).

1.2. Aplicar técnicas de minería de datos

En esta etapa, se ha utilizado el software de minería de datos Weka con el objetivo de obtener modelos de clasificación que puedan servir para predecir el rendimiento académico de los estudiantes.

En primer lugar se ha realizado un estudio de los atributos, para lo cual se ha aplicado técnicas de selección de atributos. El objetivo no es el de reducir el número total de atributos, sino el de determinar cuáles pueden ser más importantes, ya que algunos pueden estar

correlados o bien no aportar información para predecir el rendimiento final. Para ello, se han ejecutado 10 algoritmos diferentes de selección de características que proporciona Weka: CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, ReliefAttributeEval, SymmetricalUncertAttributeEval.

Atributo	Significado
NotaEntrada	Nota entrada Universidad.
Via/Opcion	Vía entrada Universidad.
Asistencia-Clases	Asistencia a clases de teoría.
Asistencia-Practicas	Asistencia a las clases de prácticas.
Expectativas-Aprobar	Expectativas por aprobar la asignatura.
Notas-Anteriores	Notas anteriores a la universidad.
NivelEconómico	Nivel económico familiar.
NivelImplicación-Profesores	Nivel de implicación de los profesores
NivelUtilización-Moodle	Nivel de utilización/uso del sistema Moodle

Tabla 5. Atributos seleccionados por todos los algoritmos de selección de características.

Tras analizar las salidas de los 10 algoritmos anteriores, se comprobó que un grupo reducido de atributos siempre era seleccionado por todos los algoritmos (ver Tabla 5).

A continuación, de entre los distintos tipos de algoritmos de clasificación que proporciona Weka, se han utilizado los siguientes:

- Basado en Bayes: BayesNet y NaiveBayes.
- Basados en Funciones: MultilayerPerceptron, RBFNetwork y SimpleLogistic.
- Basados en Reglas: JRip, NNge y PART.
- Basados en Árboles: J48 (mejora del C4.5), LADTree y RandomForest.

Se han ejecutado todos estos algoritmos utilizando los parámetros por defecto, todos los datos (con todos los atributos) y una validación cruzada de 10 iteraciones. Tras realizar todas las ejecuciones, se ha comparado la exactitud (porcentaje de estudiantes correctamente clasificados) obtenida por los algoritmos con los datos de prueba (ver Tabla 6). Se puede apreciar que en general los algoritmos obtienen una exactitud no muy alta, de entre el 53% y el 60%, y

que los algoritmos que obtienen la máxima exactitud son: NaiveBayes, PART y J48.

1.3. Analizar los modelos obtenidos

De entre los 3 algoritmos que han obtenido los mejores resultados, nos quedamos con PART y J48 al generar modelos interpretables basados en reglas y en árboles de decisión que son muy utilizados en inteligencia artificial como sistemas de apoyo a la toma de decisiones.

Algoritmo	Exactitud
BayesNet	59.2105 %
NaiveBayes	60.5263 %
MultilayerPeceptron	54.4715 %
n	
RBFNetwork	56.5789 %
SimpellLogistic	53.1579 %
JRip	53.9474 %
NNge	56.5789 %
PART	60.5263 %
J48	60.5263 %
LADTree	56.5789 %
RandomForest	56.5789 %

Tabla 6. Exactitud obtenida con los distintos algoritmos de clasificación utilizados.

En el árbol de clasificación (ver Tabla 7) se obtienen reglas de clasificación IF-THEN con sólo recorrer el árbol desde el nodo raíz a cada una los nodos terminales. Se observa cómo la variable o factor más discriminante en el rendimiento académico final es la asistencia a prácticas, que aparece en el primer nivel del árbol. Las siguientes variables más discriminantes, que aparecen en un segundo nivel del árbol, son la asistencia a las clases de teoría, el nivel de apoyo de los padres y la nota de entrada a la Universidad. Finalmente, otras variables que también influyen en la predicción del rendimiento académico y que aparecen a un nivel de profundidad mayor en el árbol de clasificación son: las expectativas por aprobar, el nivel de calidad de los recursos y materiales (NivelApoyoPadres), y el nivel de integración en clase (NivelIntegraciónEnClase). Queremos reseñar que estas dos últimas variables no habían sido seleccionadas por los algoritmos de selección de atributos.

Con respecto al modelo de clasificación obtenido por el algoritmo PART (ver Tabla 8) se observa en general que son menos reglas IF-THEN, son más cortas (menos elementos) y

afectan a menos variables o factores. Aunque casi todas las variables que aparecen ya han aparecido en el modelo anterior, hay dos nuevos factores o variables como son: el nivel de implicación de los

profesores y el nivel de dificultad/exigencia de la asignatura (NivelDificultadExigenciaAsignatura) que no habían sido seleccionados por los algoritmos de selección de atributos.

Modelo de predicción

```

AsistenciaPracticas = Baja
| AsistenciaClases = Alto: No-presentado
| AsistenciaClases = Medio: No-presentado
| AsistenciaClases = Bajo: Suspenso
AsistenciaPracticas = Media
| NivelApoyoPadres = Alto
| | NivelCalidadRecursos = Alto: No-presentado
| | NivelCalidadRecursos = Muy-alto: Aprobado
| | NivelCalidadRecursos = Muy-bajo: No-presentado
| NivelApoyoPadres = Muy-alto: Suspenso
| NivelApoyoPadres = Medio: No-presentado
| NivelApoyoPadres = Bajo: No-presentado
AsistenciaPracticas = Alta
| NotaEntrada= Alta: Aprobado
| NotaEntrada= Media
| | AsistenciaClases = Muy-alto
| | | ExpectativasAprobar = Altas: Aprobado
| | | ExpectativasAprobar = Medias
| | | NivelIntegraciónEnClase = Alto: Aprobado
| | | NivelIntegraciónEnClase = Medio
| | | NivelApoyoPadres = Muy-alto: Aprobado
| | | NivelApoyoPadres = Alto: Suspenso
| | | NivelApoyoPadres = Medio: No-presentado
| | | NivelApoyoPadres = Bajo: No-presentado
| | | NivelIntegraciónEnClase = Bajo: Suspenso
| | | NivelIntegraciónEnClase = Muy-alto: Aprobado
| | | ExpectativasAprobar = Muy-altas: Aprobado
| | | ExpectativasAprobar = Bajas: No-presentado
| | | ExpectativasAprobar = Muy-bajas: Suspenso
| | AsistenciaClases = Alto: Aprobado
| | AsistenciaClases = Medio: Suspenso
| | AsistenciaClases = Bajo: No-presentado
    
```

Tabla 7. Árbol de clasificación obtenido con el algoritmo J48.

Modelo de predicción

```

AsistenciaPracticas = Baja AND AsistenciaClases = Muy-bajo: No-presentado
ExpectativasAprobar = Muy-altas: Aprobado
AsistenciaPracticas = Alta AND NotaEntrada = Alta: Aprobado
NivelImplicaciónProfesores = Bajo AND ExpectativasAprobar = Bajas: No-presentado
ExpectativasAprobar = Medias AND NivelDificultadExigenciaAsignatura = Alta: Aprobado
ExpectativasAprobar = Bajas: Suspenso
ExpectativasAprobar = Altas: Aprobado
    
```

Tabla 8. Reglas de clasificación obtenidas con el algoritmo PART.

Conclusiones

Este trabajo muestra cómo se pueden crear modelos de predicción del rendimiento académico utilizando técnicas de minería de datos. En

concreto, se ha realizado un estudio con 280 alumnos que han cursado una asignatura de informática en el primer curso de las nuevas titulaciones de grado de la EPS en la UCO. Es

importante indicar que casi la totalidad del tiempo y esfuerzo realizado durante el desarrollo de esta experiencia se ha dedicado a la captura y pre-procesado de la información. Esto es debido a la gran cantidad de información que se quería recoger de muy distintas fuentes, además de que se ha tenido que diseñar una encuesta para capturar información.

Con respecto a la utilidad de los modelos obtenidos, los dos modelos mostrados permiten detectar a los alumnos que pueden fracasar o abandonar. Estos nos indican las variables o factores que más afectan, y más concretamente qué valores y cómo están relacionados con otras variables o factores. También queremos indicar que algunos de los atributos discriminantes (ver Tabla 5) no han aparecido en ninguno de los dos modelos obtenidos, como por ejemplo el nivel de utilización de Moodle.

Con respecto a la exactitud de los modelos, los porcentajes de clasificación obtenidos no han sido muy altos, lo que nos indica la dificultad del problema de la predicción del rendimiento académico al estar afectado por muchos factores. Creemos que algunas formas de mejorar la exactitud podrían ser:

- Utilizar sólo información sobre una única asignatura o una misma titulación, ya que hay factores que pueden variar entre una titulación y otra. Para ello se necesitaría tener más datos/alumnos de cada titulación.
- Utilizar las notas obtenidas en las siguientes convocatorias de las asignaturas durante los siguientes 2 años. De esta forma se tendrá una mayor fiabilidad sobre los alumnos que siguen sin aprobar y cuales abandonan.
- Recoger la información necesaria de otras fuentes más fiables que desde una encuesta realizada a los propios alumnos, ya que se basa en su sinceridad. Aunque variables como las notas anteriores se podrían obtener desde otras fuentes, hay variables como la situación económica familiar que son mucho más difícil de obtener.

Agradecimientos

Los autores agradecen el soporte económico proporcionado por el ministerio de educación, ciencia e innovación (Proyecto TIN-2011-22408) y la Junta de Andalucía (Proyecto P08-TIC-3720).

Referencias

- Areque, F., Roldan, C., Salguero, A. *Factors influencing university drop out rates*. Computers & Education 53, 563-574, 2009.
- Cabrera, L., Tomás, J., Álvarez, P. y Gonzalez, M. *El problema del abandono de los estudios universitarios*. RELIEVE, 12, (2) 171-203, 2006.
- De Miguel, M. y Arias, J.M. *La evaluación del rendimiento inmediato en la enseñanza universitaria*. Revista de Educación, 320, 353-377, 1999.
- España. *Ley Orgánica 15/1999, de Protección de Datos de Carácter Personal*. BOE 298, 43088-43099, 1999.
- González Tirados, R.M. *Principales dificultades en el rendimiento académico en primer año de carreras de ingeniería*. En CIDE: La investigación educativa sobre la universidad. Madrid: CIDE. 1990.
- Hernández, M. *Causas del Fracaso Escolar*. XIII Congreso de la Sociedad Española de Medicina del Adolescente. 2002.
- Mitchell, T.M. *Machine Learning*. McGraw-Hill. 1997.
- Kotsiantis, S. *Educational Data Mining: A Case Study for Predicting Dropout*. Int. J. Knowledge Engineering and Soft Data Paradigms. 1(2), 101-111, 2009.
- Romero C., Ventura. S. *Educational Data Mining: A Survey from 1995 to 2005*. Expert Systems with Applications, 33(1), 135-146, 2007.
- Rice, W.H. *Moodle E-learning Course Development. A complete guide to successful learning using Moodle*. Packt publishing. 2006.
- Saldaña, R. *Éxitos y fracasos en la Universidad*. En M. Latiesa (comp.), Demanda de Educación Superior y Rendimiento en la Universidad. Madrid: CIDE. 1986.
- Witten, I.H., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. 2005.